Data Immersion

Exercise 5.5

8/28/23

Micky Smith

Intro to Predictive Analysis

1.

You learned about linear regression in this Exercise, but you'd also like to know what logistic regression is. Conduct some research on logistic regression and explain how it differs from linear regression. When would you use logistics instead of linear regression and why?

Linear regression deals with the relationship between independent and dependent variables in regard to predictive analysis. Logistic regression deals with classification analysis where it is used to predict a binary outcome based on a set of independent variables.

You would use logistic regression when solving classification problems, using categorical variables, and when trying to find the s-curve to classify samples.

2.

Take a look at the linear regression below. It shows a relationship between the number of clients at Pig E. Bank and the number of alerts for fraudulent activity at the bank. Describe the relationship between these two variables. Based on the results, how would you assess the fitness of this model in predicting alert volume based on the number of clients?

The relationship between clients and alert volume is that if Pig E. Bank has more clients, then the alert volume is higher. From the graph we can see that the r-squared has a value of 0.8648 which provides a positive correlation between the clients and alert volume for fraudulent activity. The relationship isn't 1:1 or the r-squared value would be 1 exactly, however, the graph still provides a strong relationship between both variables.

3.

Read the scenarios below, then decide which predictive model you'd use in each one. Provide a short explanation for the rationale behind your decisions.

- **Scenario A: As an analyst for a large financial institution, your job is to perform research and develop models that predict the future values of previous metals. You theorize that the global oil price can be predicted based on the unemployment rates of the top 20 counties in GDP. Would you use a regression model or classification model to validate your theory? What specific algorithm would you use for this predictive model and why?**

I would use a linear regression model in this scenario. We are working with two quantitative variables. Unemployment rates would be the independent variable (x-axis), while the oil price would be the dependent variable (y-axis). We would be looking for the best fit line within the relationship between these two variables.

- **Scenario B: You're a data analyst for an online movie provider that collects data on its customers' viewing habits. Part of your job is to support the company's efforts to display movies that customers are likely to enjoy prominently on their profile page and keep the movies they're least likely to enjoy off their profile page altogether. To this end, your company has asked you to predict which customers are most likely to watch a romantic comedy starring Adam Sandler and Drew Barrymore. Would you use a regression or classification model for this? What specific algorithm would you use and why?**

I would use a classification model because this scenario deals with categorical variables. We are predicting a binary output of customers who will watch a romantic comedy with Adam Sandler and Drew Barrymore vs those who will not. We would use a random forest tree which would be made of several different decision trees based on the qualitative data to determine whether a customer would like the movie or not.

4.

Imagine you were involved in collecting the data that was used in the linear regression in step 2. What types of bias could have arisen when collecting the data and why?

The potential for bias within the linear regression in step 2 is that it only provides a small subset of the data, which can provide misleading data by providing a lack of results and provide a skewed result of the data.

There is a collection bias because we are unsure where the data was collected and why only a random sample is being used. We would need further details into the data to be able to ensure it's accuracy.